DOCUMENT RESUME

ED 118 601 TM 005 085

AUTHOR Thorndike, Robert M.; Waldram, Carol A.
TITLE Content Validity and the Uses of Student Ratings.
NOTE 13p.

EDRS PRICE MF-$0.83 HC-$1.67 Plus Postage
DESCRIPTORS Evaluation Needs; Higher Education; *Rating Scales; *Student Evaluation of Teacher Performance; *Student Opinion; *Teacher Evaluation; *Test Validity

ABSTRACT

The validity of a rating instrument depends not only upon the content of the instrument and the purposes the instrument is to serve, but also upon its relationships to criteria external to the instrument. Formative and summative evaluation require two distinct types of instruments which yield different types of information. The problem that remains in the development of formative rating devices is a definition and enumeration of the domain of teaching behaviors. A solution to this problem possibly would be different for various subject areas and teaching styles. Careful sampling of this domain may require a set of rating devices which are hierachically arranged. Current rating instruments represent a step toward the development of broad-band devices, which could be used at the highest level of the hierarchy to identify problem areas, and may also be useful for summative evaluation purposes. Evaluation instruments yield information that is reliable and logically expected to be related to quality of teaching performance. What is lacking is evidence relating ratings to important educational outcomes. (Author/RC)

****
* Documents acquired by ERIC include many informal unpublished *
* materials not available from other sources. ERIC makes every effort *
* to obtain the best copy available. Nevertheless, items of marginal *
* reproducibility are often encountered and this affects the quality *
* of the microfiche and hardcopy reproductions ERIC makes available *
* via the ERIC Document Reproduction Service (EDRS). EDRS is not *
* responsible for the quality of the original document. Reproductions *
* supplied by EDRS are the best that can be made from the original. *
****

ED118601

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PEPSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

Content Validity and the Uses of Student Ratings

Robert L. Thorndike and Carol A. Waldram

Western Washington State College

Numerous methods of evaluating teaching have arisen in an effort to assess the effectiveness of instructors in their classrooms. Colleague, administrator, student and self ratings are just a few of the methods in use. A survey conducted by the American Council on Education in 1966 revealed 15 sources of information commonly used to assess teaching effectiveness (Astin and Lee, 1967). A total of 1,110 universities, colleges of education, liberal arts colleges and junior colleges responded to the questionnaire. Sources of information by order of reported usage were as follows: chairman evaluation, dean evaluation, colleagues' opinions, scholarly research and publications, informal student opinions, grade distributions, course syllabi and examinations, committee evaluation, student examination performance, self-evaluation or report, classroom visits, systematic student ratings, enrollment in elective courses, long-term follow-up of students, and alumni opinions.

An examination of the results of this survey reveals a relatively infrequent use of systematic student ratings. The percentage of institutions reporting their use in all or most departments varied from a high of 26.5% in university colleges of agriculture to a low of 4.9% in teachers colleges. Thirty one percent of all the institutions responding to the questionnaire, however, reported using either informal student polls or other devices based on student ratings for the evaluation of teaching effectiveness (Kent, 1967).

Although there is no comparable data for 1975, an increased emphasis on student input and demands for accountability support the assumption that systematic student evaluation of instructors has been used increasingly since the 1966 survey. Student activism in the late sixties and in 1970 brought public attention to students' discontent with the status quo. Accompanying this movementwere demands for relevant curricula and a greater voice in the educational process. Demands for instructor accountability and assessment programs have also come from educators themselves (see Popham, 1974, for example). These demands have generally focused on a need for an improvement in the quality of education at the university level or a desire to take teaching effectiveness into consideration when requests for tenure and promotion are made. Whatever the purposes of evaluation programs, two factors support the increased use of student input. The involvement of students by means of an evaluation procedure to some extent may appease student demands for a voice in the selection of instructors for tenure and promotion. Secondly, students provide a convenient source of information. Administrators and colleagues are frequently too busy with their own duties to spend the time necessary in a classroom observing an instructor.

The use of students in evaluation programs is defensible on grounds other than the appeasement of student demands or as a matter of convenience. Logically,

TM005 085

students should provide a primary source of information in an assessment program. The student rating an instructor at the end of a quarter or semester generally has spent 30 to 60 hours observing and interacting with the instructor in the classroom. This is a source of information on the habitual classroom behavior of faculty which should not be wasted.

Uses of Student Evaluations

Student ratings of faculty effectiveness are commonly put to two uses: faculty evaluation and faculty improvement. Faculty and administrative committees frequently require that an instructor support requests for promotion and/or tenure by submitting results from student evaluations. Evaluations collected for institutional decision making purposes are analogous to evaluation procedures which Bloom, Hastings, and Madaus (1971) have labelled "summative." A second purpose of evaluations of teaching ability is to provide the instructor with information on the success of his teaching efforts. Feedback can be a useful means of helping the instructor to change his teaching behaviors. This use of evaluations is similar to Bloom's category of "formative" evaluation in that it is designed to pinpoint the part of a task not mastered.

Summative evaluation seeks an overall assessment of the ability or performance of the instructor as compared to other instructors, whereas formative evaluation is diagnostic and requires precise information about specific areas of competence and weakness. Summative evaluations require global ratings similar to those currently in widespread use. When the goal of the evaluation procedure is to help the instructor improve his performance, a precise description of his current teaching behavior and specific suggestions for change are necessary.

Clearly, an evaluation procedure suited for one purpose does not adequately function in the other capacity. Two types of instruments are necessary. Formative evaluation of teaching requires a definition of the dimensions of teaching behaviors and their effects on the learning process. An enumeration of the behaviors to be included in such an evaluation procedure will most likely be different for various subject areas and teaching styles. A complex catalog of behaviors would result, and rating could become a time-consuming and difficult task. Institutions desiring to use formative evaluations could overcome this problem by the use of a set of rating devices which are hierarchically arranged. A first level rating device would be one in which broad categories of behaviors are rated. Many current evaluation instruments potentially fall into this class of evaluations. Problem areas identified from the first level instrument would be investigated further by means of a second level instrument dealing with only those teaching behaviors related to the problem.

There are two advantages to this sequential approach. First, it will be more economical of time and material because ratings of specific behavior using second level instruments need only be obtained for those classes of behaviors which are identified as problem areas by the first level instrument. Second, a hierarchical set of instruments is more flexible. Optimum teaching behaviors are probably different for different disciplines and for different teaching objectives. A hierarchical organization of rating instruments would permit different rating scales to be used to fit course content and the cognitive objectives of the particular instructor.

Desirable Properties of Rating Scales

One of the first requirements of an evaluation device is that it yield consistent information across reasonable time intervals. It is a logical and statistical fact that a measure cannot be valid unless it is reliable. If student ratings of faculty effectiveness are reliable, then ratings should be consistent across time when the instructor does not know the results of the ratings. Finding stability in ratings under these conditions implies that the instructor's behavior is consistent and that students' perceptions and ratings are also consistent.

Evidence that students can provide reliable ratings of faculty is available from numerous studies (Costin, 1968, 1971; Lovell and Haner, 1954), including three conducted at Western Washington State College. One study examined the reliability of two different response formats (Likert and Thurstone) for a 40-item experimental rating scale (Thorndike, Easton, and Scheehe, 1973). In the Likert format students were required to make quantitative judgments, whereas in the Thurstone format students merely chose items that were descriptive of the instructor and course. Over a one week interval the median reliability for the Likert form was about .70, and the Thurstone format yielded a reliability of about .75 for individual responses from six different classes. A second study, using the Thurstone form of this instrument, obtained one week retest reliabilities from 25 classes in five departments in which the average individual response reliability was .84 (Easton, 1973). The third study found similar reliabilities for an approximate two week retest interval using 16 classes in five different departments (Waldram, 1974). This study used the Likert form of the experimental questionnaire and the 11-item scale used for summative evaluation by the institution. The results of these studies with different response formats, items, various classes from numerous departments, and different time intervals suggest that students are capable of providing stable ratings of faculty performance.

A more pressing concern in the evaluation of faculty performance by students is not whether the individual student is consistent in his judgments, but rather whether the conbined judgments of a class yield a stable profile of the individual instructor. Profiles of mean ratings are generally the information reported to the faculty member. Reliability of group mean profiles were calculated in the first two of the above studies conducted at Western Washington State College. In every class, the profile reliability was greater than .90, and in most classes it was .95 or higher. This indicates an extremely high stability of judgments, leading one to have confidence in the accuracy of profiles obtained from student ratings.

Evaluation instruments thus yield information that is reliable. To be useful, however, the information obtained from the instrument must also be valid. Certainly the relevance of ratings obtained for either formative or summative purposes is to some extent defined by the questions asked the students. Many evaluation forms ask the student to rate only the presence or absence of desirable behaviors. Behaviors which are characteristic of poor teaching are not commonly presented. It seems inappropriate to assume that the absence of a behavior characteristic of good teaching implies the presence of an undesirable

behavior. This is particularly important in the use of evaluations for formative purposes where a precise description of the instructor's behavior is needed. At most the student can accurately rate only those behaviors that he can observe. Examples of items that fall outside of this category are easy to find on many evaluation questionnaires. Clearly, ratings of "demonstrated a sincere interest in the subject" or "responded to questions to the best of his ability" are not based on observable behaviors. Many students may not even be able to respond to such questions as "How available is the instructor outside of the classroom?" Ratings of the importance and usefulness of course content are bound to vary greatly with the immediate applicability of the material covered and course level.

With the exception of items describing unobservable behaviors, student evaluations "look as if" they measure at least some of the dimensions of teacher effectiveness. Students can logically rate behaviors such as the apparent organization of course materials and lecture presentations, the difficulty and speed at which the course material was mastered, the use of teaching tools, and so forth. Such observable behaviors make up the largest part of most evaluation devices. This type of examination of content of instruments falls under the heading of face validity. The development of a teacher evaluation instrument generally includes this type of validation when the questionnaire is submitted to faculty members for approval.

This face validity is not sufficient. Evaluation instruments should meet this standard if students are to take the evaluation procedure seriously and if teachers are to attend to the results of ratings, but this standard is far from complete. Sockloff (1973) in an examination of around 500 rating instruments found that the typical instrument contains 11 to 30 items and was derived largely from other instruments. Few rating form authors have demonstrated content validity based on a thorough task analysis in which the many elements of the teacher's job are enumerated and covered by the instrument. Even instruments developed from general objectives, lists of "critical behaviors" or from other attempts at covering the domain of good teaching cannot be certified as possessing content validity. Indeed, a rating device which is content valid for one teacher or discipline may not be satisfactory for another due to different subject matter or method of presentation. The purpose of the evaluation procedure (summative or formative) is also an important consideration regarding the adequacy with which important teaching behaviors have been sampled. As was noted previously an evaluation procedure suited for summative purposes does not function adequately in the other capacity or vice versa. These factors support the development of multiple scales which might profitably be hierarchical in organization.

Current evaluation instruments operationally define the good teacher as one who is well organized, explains grading procedures, gives assignments and tests that the students like, keeps frequent office hours and so forth. Such is the picture of the good teacher which is provided by many of the so called face valid measures of teacher effectiveness in common use today. Unfortunately this description provides a circular definition of good teaching by defining it as a process. Some institutions have had some success in using these ratings in conjunction with counseling designed to improve teaching (Aleamoni, 1973; Tuckman, 1973). This work demonstrates that faculty can alter their behavior so that higher student ratings are obtained. One important question remains: Is

the instructor a better teacher as a result of this alteration in behavior? An escape from this circularity of definition is possible not by defining good teaching as a process, but instead defining it as a product. This approach to the validation of student evaluation instruments calls for a demonstrated relationship to criteria external to the instrument.

The definition and measurement of such criteria for empirical validity studies are difficult. Evidence which links student progress toward the achievement of course objectives to their ratings of instructor effectiveness is a vital step toward establishing the validity of student evaluation procedures. However, some of the most important objectives of higher education are difficult to quantify. The goals of educators are frequently not just limited to subject-matter achievement. Educators commonly seek affective changes as well as cognitive changes in students. An affective objective in a course on ethnic studies, for example, may be a change in attitude toward a minority group; an affective objective for an art appreciation class might be the aesthetic enjoyment of masterpieces of a given period in art history. Measures of the attainment of cognitive objectives, therefore, only partially measure teacher effectiveness. However, they are typically the only ones systematically evaluated by faculty.

Other problems are encountered when one tries to measure cognitive objectives. Correlations between individual scores on measures of cognitive objectives and an instructor's ratings give little validation information. Evaluations completed for summative or formative purposes should discriminate between better and poorer instructors not better and poorer students. Evidence for the empirical validity of these evaluations must come from comparisons among instructors and not from comparisons among students rating the same instructor. Traditional measures of the amount students learn from a course, such as test scores and grades, are not adequate measures of the amount of progress toward cognitive objectives across classes (or within a class for that matter). When student progress toward the achievement of course objectives is to be used as the criterion measure, all relevant variables other than the effects of the teaching itself should be held constant. If grades or test scores are used as the criterion measure, an obtained relationship may be due to grading policies or variations in student ability across classes rather than to the effects of the teaching itself. An objective measure of final achievement in a course with some correction for initial ability provides a more satisfactory criterion measure.

Validity Studies

Several studies have compared student evaluations of instructors to measures of amount learned from a course. The Rodin and Rodin study conducted in 1972 is a well known example. These authors found a negative relationship (-.75) between ratings of instruction and the grades received in an undergraduate calculus class. Course grades were assigned on the basis of the number of criterion problems passed. Classes were made comparable by partialing out initial ability as measured by grades from the previous quarter. Rodin and Rodin concluded that the resulting negative relationship between student evaluations and grades indicated a negative relationship between amount learned and student evaluations.

This conclusion is true only to the extent that grades were a measure of amount learned by the students in the classes. Grades provide only five levels

of amount learned from a course. Students were allowed to attempt a given type of criterion problem up to six times, and the final grades for the course were assigned on the basis of the number of criterion problems completed without error.

The Rodin and Rodin study has been criticized for the use of a one-item question evaluation procedure and the structure of the classes participating in the study (Frey, 1973; Gressner, 1973). Students rated the instructor on the following question: "What grade would you assign to his total teaching performance?" Such one-item methods of evaluating teacher effectiveness involve only one global dimension of teaching and may well be unreliable due to the shortness of the form. Students enrolled in the course met for a large lecture 3 days per week and for a small recitation section 2 days per week with one of eleven graduate teaching assistants. Teaching assistants devoted 1 day per week to answering questions about homework and lectures. Administration of test problems and reviews of preceding test problems took up the other day each week. It was the graduate teaching assistants that the students rated. Frey (1973) has commented that "although Rodin and Rodin purported to investigate their students' ability to identify good teachers their research assessed the effectiveness of graduate teaching assistants in complementing the teaching style of the authors (p. 83)."

Frey (1973) replicated the Rodins' study with several modifications. Student ratings of instruction in undergraduate calculus classes were correlated with class performance on a common final examination. The rating instrument consisted of 18 items which were scored on a 7-point scale. Students were sent these forms early in the quarter following completion of the course. The returned forms (75%) from the classes were factor analyzed. The items were grouped into 6 factors as a result of this analysis. Regressed final examination scores were calculated for each student by computing the difference between his observed score and the score predicted on the basis of his Scholastic Aptitude Test scores and adding this to the average observed score of all the students taking the examination. The average regressed final examination score for each class was correlated with the average ratings on items with each factor for each instructor. The obtained correlations were positive in all cases.

It is difficult to generalize from this study for several reasons. A sample bias may have resulted from ratings being collected after students received their final grades for the course. While the authors did look at possible biases in ratings due to the effects of grades or the selective return of questionnaires, no control is possible at the instructor level for overall changes in evaluations. Thes authors also utilized a more refined analysis of the evaluation form than the data warrented. It seems unlikely that the 18 items on the evaluation instrument yielded 6 <u>independent</u> sources (factors) of information in that at least four of the factors correlated highly with the same criterion variable, suggesting that they were at least somewhat redundant.

Other studies have approached the problem of controlling for initial abilities in a number of ways. Some studies have attempted to control for initial ability by using the same students taking a course from a number of different instructors. Others have randomized the assignment of students to classes.

Gressner (1973) attempted to circumvent the problem of controlling for initial ability by using sophomore medical students taking a basic science class. Ten different instructors taught different subject areas of the course. Level of student achievement was measured by performance on the National Medical Board Examination. Students rated each of the 23 subject areas covered in the course on a 3-point scale for content and organization, and for presentation. The correlation between scores on sections of the National Medical Board Examination pertaining to material covered in one of the subject areas of the course and the two measures of teaching effectiveness were .77 and .69, respectively. This use of the same students does not represent an effective control procedure. The obtained correlations may have been due to exposure to the subject areas prior to the course in question. Students may have rated higher those areas with which they were most familiar.

Sullivan and Skanes (1974) randomly assigned students to different class sections of 10 first-year courses in biology, chemistry, mathematics, physics, science, and psychology. Mean ratings for each class on a question dealing with the overall competence of the instructor were correlated with average class final examination scores on a common final examination for each course. Of the 10 obtained correlations, 9 were positive, 8 were above .32, and the average correlation was .39. Subsequent analysis of experienced, inexperienced, and graduate student instructors of the 40 class sections of the introductory psychology course revealed that correlations were higher for experienced faculty members ($r=.69$; $p<.01$) than for inexperienced faculty members ($r=.13$; $p>.05$) and the lowest for the 13 graduate student instructors ($r=.01$; $p>.05$). Sullivan and Skanes concluded that under certain circumstances student ratings are valid, and that this may in part explain the contradictory results obtained in the Rodin and Rodin, and Gressner studies. Positive correlations between ratings and achievement measures are more likely to be obtained from full-time experienced instructors than from inexperienced or part time (graduate) instructors.

While the random assignment of subjects to class sections in the Sullivan and Skanes study does not guarantee an equalization of the effects of varying initial abilities on final achievement test scores, it at least is a step towards its elimination as a source of bias in experimental outcomes. Ratings in this study as in the Rodin and Rodin and Gressner studies, however, were based on 1 or 2 item evaluation instruments of unknown reliability.

Students entering a specific course often do not have homogeneous backgrounds (as in the Rodin and Rodin, 1972 study), and it is usually impractical to randomly assign students to class sections. Another method has been used to assess the amount learned in a course. This was done by measuring student progress as determined by the difference in scores on a test at the beginning of a course and at the end of the course. Bimson (1937) used this method to determine student progress in 25 high school beginning algebra, general science and American history classes. Odd-numbered questions from standardized tests were given as a preliminary test at the beginning of the semester, and the even-numbered items from the same test were given at the end of the semester. Average student progress, as measured by the difference between these scores, was compared to supervisors' judgments of teaching ability. The median ranks for instructors were determined, and teachers with a higher than median rank were considered

"good" teachers. The results from this study demonstrated a general tendency for students with "good" teachers to make somewhat greater progress than students with "poor" instructors; however, statistical significance cannot be determined from the data reported. Bimson's use of gain scores to measure student progress controls for initial knowledge and ability to a greater extent than methods of assessing achievement used by the other authors cited.

The gain score approach to the measurement of amount learned was adapted by Morsh, Burgess and Smith (1956) in the assessment of the validity of student evaluations of instructors in an Air Force mechanics course. The matching of students entering various classes was impossible. Classes were matched statistically by means of a 65-item pretest administered to all students before they entered an 8-day hydraulics course. The pretest was constructed to measure general knowledge in the subject matter to be taught in the course. After completion of the course all students rated their instructor on overall competence and four other qualities. Students then took a post-test over the subject-matter covered in the course and a job-sample type performance test. Instructor effectiveness was evaluated by means of average student post-test scores adjusted for differences in average pretest scores. These authors report making adjustments via a multiple regression equation computed from within instructor and between class variances and covariances. It is unclear from the information provided, however, just how this was acomplished. Students' overall rating of the instructors correlated significantly with measures of performance ($r=.39$; $p<.05$) and written gains ($r=.32$; $p<.05$). Ratings of "teaching ability" were correlated .42 and .41 with performance and written criteria, respectively.

Bimson (1937) and Morsh, Burgess and Smith (1956) present a technique which is feasible for use in the empirical validation of teaching evaluations in the college or university setting. This is illustrated by two studies in which the amount learned in a course was defined by the average gain in a class as measured by a pre and post course test. A total of 102 students (52% of those enrolled) attending 3 class sections of a psychology course dealing with the application of psychological principles to teaching served as subjects.[1] The procedure was repeated on a sample of 319 students (59% of those enrolled) attending 4 class sections of an introductory psychology class. Two forms were used for student evaluation of faculty. These forms included the standard evaluation form in use at Western Washington State College and an experimental form.[2] The reliability of both forms has been demonstrated (Thorndike, Easton, and Sheehe, 1973; Waldram, 1974). The standard survey consists of 12 positively worded statements. Ten of the statements describe teaching behavior and the last two items are general ratings of the instructor and course. Forty statements describing positive to negative teaching behaviors are included on the experimental survey.

---

[1] It is common practice in the literature to not report the proportion of students enrolled in a course who actually completed the course evaluations. Such reports should not be omitted. In this study, for example, it was found that in 2 of the 3 classes students participating in the experiment obtained significantly higher grades than those who did not.

[2] Copies of the evaluation forms may be obtained from the senior author.

Only the experimental form was used in rating instructors teaching introductory psychology.[3]

A means of accounting for the amount of knowledge gained from a course was necessary. This was done by administering a test during the first class session of the course and the same test after the course was completed. Instructors teaching the course in question submitted a number of 4-alternative, multiple choice questions which they felt reflected what they wanted the students to learn during the quarter. All questions were pooled and submitted to each instructor. Working independently, each professor checked off the items which he believed measured his course objectives. A test was then constructed using only those questions that all instructors checked. This process resulted in a 32-item test in the original study, and a 30-item test which was administered to the introductory psychology classes. To prevent instructors from teaching to the test, participating instructors did not have access to the test before the final week of classes.

Both evaluation forms were administered simultaneously during the last two weeks of the quarter. The order of presentation of forms was counterbalanced. All students remained anonymous until after the evaluations were completed. They were then asked to submit their name or student number; with few exceptions all students complied.

The average gain score for each class was compared to the average instructor rating. One faculty member was rated significantly higher than either of the others in the educational psychology course, and his class showed significantly higher gain scores than either of the other two classes. The two lower rated instructors showed a similar ordering in gain scores and ratings. Ratings of the introductory psychology instructors ranged from 2.86 to 4.08, and gains ranged from 3.70 to 9.65. A similar ordering of ratings and gain scores was found in this course with one exception. While the average rating of the instructor rated third exceeded the rating of the lowest rated instructor (3.77 as opposed to 2.86), the gain score in this class did not (3.70 as opposed to 6.47). Subsequent analysis of the course content of the introductory psychology classes indicated that this may have been a function of the relative emphasis placed on subject matter. All instructors, except the instructor who was rated third, approached the teaching of the class as a general survey course; however, the third rated instructor emphasized physiological psychology. The index of amount learned in this sample, consequently was not representative of the course content in this class.

An expansion of this research technique to courses which include more class sections may provide much needed information as to the relationship between rating instruments and important educational outcomes. Many courses at institutions larger than Western Washington State College provide an ideal format for such research. Courses with many class sections taught during the same

---

[3] This was due to a limited access to the official school form and an obtained correlation of .71 between average ratings on the experimental and standard forms.

quarter and with a common final examination can easily be used.

In summary, the validity of a rating instrument depends not only upon the content of the instrument and the purposes the instrument is to serve, but also upon its relationship to criteria external to the instrument. Formative and summative evaluation require two distinct types of instruments which yield different types of information. The problem that remains in the development of formative rating devices is a definition and enumeration of the domain of teaching behaviors. A solution to this problem possibly would be different for various subject areas and teaching styles. Careful sampling of this domain may require a set of rating devices which are hierarchically arranged. Current rating instruments represent a step toward the development of broad-band devices which could be used at the highest level of the hierarchy to identify problem areas and may also be useful for summative evaluation purposes. Evaluation instruments yield information that is reliable and logically expected to be related to quality of teaching performance. What is lacking is evidence relating ratings to important educational outcomes.

References

Aleamoni, L. The usefulness of student evaluations in improving college teaching. In A. Sockloff (Ed.), Proceedings: Faculty Effectiveness as Evaluated by Students, Philadelphia: Measurement and Research Center of Temple University, 1973, 42-58.

Astin, A. W., & Lee, C. B. T. Current practices in the evaluation and training of college teachers. In C. B. T. Lee (Ed.), Improving College Teaching, Washington D. C.: American Council on Education, 1967.

Bimson, O. H. Do "good" teachers produce "good" results? The North Central Association Quarterly, 1937, 12, 271.

Bloom, B. S., Hastings, J. T., & Madaus, G. F. Handbook on Formative and Summative Evaluation of Student Learning. New York: Mcgraw-Hill Book Company, 1971.

Costin, F. A graduate course in the teaching of psychology: Description and evaluation. Journal of Teacher Education, 1968, 19, 425-432.

Costin, F. An empirical test of the "teacher centered" versus "student centered" dichotomy. Journal of Educational Psychology, 1971, 63, 410.

Easton, J. Q. An examination of the generalizability of a teacher rating form. Unpublished master's thesis, Western Washington State College, 1973.

Frey, P. W. Student ratings of teaching: Validity of several rating factors. Science, 1973, 182, 83-85.

Gressner, P. K. Evaluation of instruction. Science, 1973, 180, 566-559.

Kent, L. Student evaluation of teaching. In C. B. T. Lee (Ed.), Improving College Teaching, Washington D. C.: American Council on Education, 1967.

Lovell, G., & Haner, C. Forced-choice applied to college faculty rating. Educational and Psychological Measurement, 1955, 15, 291-304.

Morsh, J. E., & Burgess, G. G., & Smith, P. N. Student achievement as a measure of instructor effectiveness. Journal of Educational Psychology, 1956, 47, 79-88.

Popham, J. W. Higher education's commitment to instructional improvement programs. Educational Researcher, 1974, 3 (11), 11-13.

Rodin, M., & Rodin, B. Student evaluations of teachers. Science, 1972, 177, 1164-1166.

Sockloff, A. L. Instruments for student evaluation of faculty: Ideal and actual. In A. Sockloff (Ed.), Proceedings: Faculty Effectiveness as Evaluated by Students, Philadelphia: Measurement and Research Center of Temple University, 1973, 132-151.

Sullivan, A. M., & Skanes, G. R. Validity of student evaluation of teaching and the characteristics of successful instructors. Journal of Educational Psychology, 1974, 66, 534-590.

Tuckman, B. W. A system for helping teachers to change their affective behavior through feedback. In A. Sockloff (Ed.), Proceedings: Faculty Effectiveness as Evaluated by Students, Philadelphia: Measurement and Research Center

Throndike, R. M., Easton, J., & Sheehe, C. Reliability of Likert and Thurstone scales. Paper presented at the annual meetings of the Psychometric Society, Chicago, March 1973.

Waldram, C. A. The reliability and validity of student evaluation of teaching at Western Washington State College. Unpublished honors thesis, Western Washington State College, 1974.